DOCUMENT RESUME

ED 057 084 TM 000 927

AUTHOR Komulainen, Erkki
TITLE Investigations into the Instructional Process II. Objectivity of Coding in a Modified Flanders Interaction Analysis.
INSTITUTION Helsinki Univ. (Finland). Inst. of Education.
REPORT NO RB-27
PUB DATE Dec 70
NOTE 31p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Classification; *Classroom Observation Techniques; *Codification; Content Analysis; Courses; *Instruction; *Interaction Process Analysis; Measurement Instruments; *Reliability; Video Tape Recordings
IDENTIFIERS Finland; *Flanders Interaction Analysis; Helsinki

ABSTRACT

The reliability of coding problems associated with observation studies is discussed. The purpose is two-fold: a) to examine coding reliability by applying the profile method to two coding occasions separated by a lengthy time interval with the object of determining both within-occasion reliability (agreement) and between-occasion reliability (constancy); and b) to develop a method for the measurement of the reliability of any one individual category. (MS)

RESEARCH BULLETIN

Institute of Education
University of Helsinki

Head:
Matti Koskenniemi
Professor of Education

Snellmaninkatu 10 A
Helsinki 17
Finland

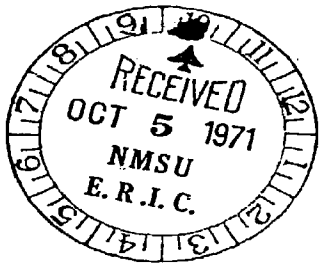
No. 27

December 1970

Erkki Komulainen

INVESTIGATIONS INTO THE INSTRUCTIONAL PROCESS

II. Objectivity of Coding in a Modified Flanders Interaction Analysis



Erkki Komulainen

INVESTIGATIONS INTO THE INSTRUCTIONAL PROCESS

II. Objectivity of Coding in a Modified Flanders Interaction Analysis

Institute of Education
University of Helsinki
1970

Contents:

Investigations into the Instructional Process

II Objectivity of Coding in a Modified Flanders Interaction Analysis

1. Introduction

This report deals with the reliability of coding problem associated with observation studies. The repeatability of videotaped situations adds new features to such analysis. The intention is:

1) to examine coding reliability by applying the customary profile method (Flanders 1965, 23-30) to two coding occasions separated by a lengthy time interval, with the object of determining both within-occasion reliability (agreement) and between-occasion reliability (constancy). The coefficients obtained will be considered according to

   a) school subjects,
   b) coder pairs and
   c) coding occasions;

2) to develop a method for the measurement of the reliability of any one individual category and to consider the coefficients obtained according to

   a) school subjects,
   b) coder pairs,
   c) coding occasions and
   d) the order of coding.

The present study is a sequel to a previous report published in this series (Koskenniemi & Komulainen 1969). The same material was dealt with in both studies. The videotaping took place in the laboratory class of the University of Helsinki Institute of Education. However, the reliability analysis only related to a total of 10 lessons in four different subjects.

Table 1. Material of the study

Subject	Date of video-taping	1st coding T1	2nd coding T2	T1/T2 comparison
1. Civics	Oct. 30	X	X	X
2. Civics	Nov. 27	X		
3. Arithmetic	Nov. 14	X	X	X
4. Arithmetic	Nov. 20	X		
5. Religion	Oct. 25	X	X	X
6. Religion	Oct. 28	X		
7. Religion	Nov. 1	X		
8. Religion	Nov. 8	X		
9. Finnish	Nov. 14	X	X	X
10. Finnish	Nov. 22	X		

The videotaping was carried out during the autumn term 1967. The interval between codings T1 and T2 was about three months.

The observation instrument used in the study was a 13-category Flanders-modification devised by the writer (Appendix 1).

2. On the Reliability Concept in Observation Studies

As in other measurements, in observation studies it is imperative to ensure that the measuring instrument is resistant to chance influences. In such studies the measuring instrument cannot be considered to consist in the system of categories alone but, instead, in the whole comprising both the category system and its user, and it is the reliability of this whole that is concerned. Thus, in observation studies the reliability concept has a content partly different from the one it has in other kinds of measurement (Stukát 1966, 120).

Conversion of the content of the instructional process into a form capable of quantitative treatment is called coding. Three steps can be distinguished in the coding of an interactional process (Guetzkow 1950, 47): (1) unitizing, which means the division of the sequence of events into elements in accordance with a rule agreed on in advance; (2) categorizing, which means the placement of each unit into a classification system designed in advance; and (3) attributing, which means the identification of the originator of a behaviour unit and the target of speech or any other sort of behaviour.

The present study is exclusively concerned with the reliability of categorizing, since the other two steps of coding can be considered to take place completely reliably. In the method used here, the unit was not a so-called natural unit but a time unit. Unitizing is not carried out by the coder but by a seconds counter. As a large-size seconds counter provides a frame of reference common to all coders, all the observers will perform the unitizing in the same way. Attributing, again, is already contained in categorizing, since the category employed also indicates which one of the two possible originators - the teacher or the pupil - is in question. Thus it is justifiable to maintain that examination of the reliability of the method employed has to do with categorizing alone.

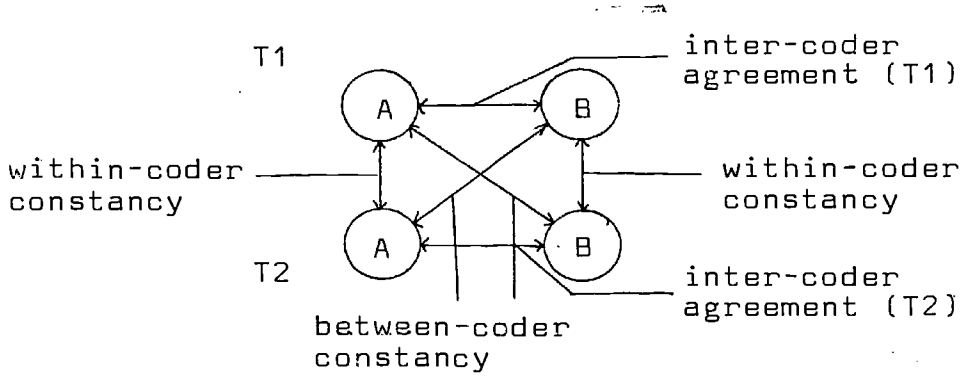
Observation reliability is most frequently defined as the degree of consistence between the results of categorizing performed by two observers simultaneously but independently. In connection with judgments, this reliability concept is usually referred to as "multi-judge" reliability (Kogan & Hunt 1950). Here, however, the term inter-coder agreement, which is the standard expression in content analysis, will be used to emphasize the objective and mechanical nature of observation, in contradistinction to the subjective element inherent in judgments. Inter-coder agreement is the similarity between the codings performed by two independent observers at the point of time T1.

The requirement has been advanced, mainly in content analytical studies (Berelson 1954), that the coder must be able to employ the category system in his codings in the same way at different times. Examination of this point has previously been possible only with literary material, and therefore it has not been investigated in the context of observation (Borgatta & Bales 1953, 566-569). Brown & Webb write: "Within-observer reliability would seem a far more useful concept than between-observer reliability for establishing reliability estimates for systematic observation" (1968, 37). Re-categorizing from a videotape and comparison of various codings done by the same person yields a reliability indicator called within-coder constancy. Also, agreement between codings of the same situation performed by different coders at different points of time can be examined. This reliability concept is termed between-coder constancy.

Certain investigators who have used the observation technique have spoken of reliability in a very broad sense. By the reliability of observation they have meant the correspondence between scores given by different observers in observation situations at different point of time. That observations are made at different points of time means that they relate either to different lessons of the same teacher or to lessons in the same

subject by different teachers (Medley & Mitzel 1963, 253-254 and 309-312). The definition involved here is based on the presumably high constancy of the trait to be measured. The high reliability postulated in this definition presupposes that observation pertains mainly to those permanent features about class-room work that are due to the teacher. Changes in class-room work from one situation to another are regarded as perplexing, and attempts are made to eliminate their influence by carrying out several observations of the same class in different situations and by making use of the average of these. Medley & Mitzel's definition ascribes to the error variance that particular feature which is of central interest in the present study: the fact that systematic differences occur in the activity of one and the same class in different situations. The reliability problem here concerned is not related to the permanence of various features but to the dependability of the measurement of the features.

The time interval between the codings was about three months. On both occasions, four observers coded the same lesson simultaneously but independently from each other. The following simplified schematic representation of the two-observer case indicates how the various agreement indices are formed.



3. The Overall Reliability as Estimated from the Marginal Distributions

The estimation of reliability from the class frequencies obtained rests on the underlying assumption that, if two observers have assessed the same number of units to a given category, they have assigned to it the same units and, thus, show inter-coder agreement. The tenability of this assumption will be discussed in chapter 5. The similarity between the frequency profiles obtained by two observers of the same lesson - i.e., profiles reflecting how the categorization system was employed by these observers - indicates the degree of agreement between them. Bales used chi-square as an index (Bales 1950, 110). It has later been realized that serious shortcomings attach to the use of both chi-square and the contingency coefficient (Flanders 1965, 30-31; Cohen 1960, 38-39), and these have been replaced by Scott's "pi" coefficient. Use is made thereby of a profile converted to percentages (to reduce the apparent disagreement arising from differences in the tempo of scoring), and an attempt is also made to estimate the amount of agreement due to chance (Scott 1955, 321-325). The coefficient is obtained from the formula

(1) $$\Pi = \frac{Po - Pe}{1.00 - Pe}$$

where: Po = observed percentage agreement
Pe = percentage agreement to be expected on the basis of chance, as obtained from (2)

(2) $$Pe = \Sigma Pi^2$$

where: Pi = the proportion of the entire sample that falls in the i:th category.

Scott's pi takes into account the fact that the agreement to be expected on the basis of chance does not equal the theoretical expectation (1/k, where k = the number of categories) but varies according to the relative frequency of occurrence of each category (Pe) in the sample to be analysed. Regarding the interpretation of this coefficient, Scott states that it roughly indicates the extent to which the coding reliability exceeds chance. The range of variation of the pi coefficient has properties similar to those of the coefficient of correlation (Cohen 1960, 41-43).

4. Results Concerning Overall Reliability

When a picture is formed of the reliability of observation, it is imperative for us, in principle, to estimate the part played by chance. More important than to test a null hypothesis is, however, to examine the size of the coefficients under varied conditions, since in practice the degree of agreement invariably exceeds chance, whatever the directions given for categorizing (Schutz 1952, 120). What is essential is to find out how far the observed reliability meets the reliability requirements the investigator has set for the problem under study. In an intensive study the reliability must be comparatively high.

In the present study the mean of all the agreement coefficients computed was .79 (Table 2).

The differences between the school subjects were not statistically significant (t -test). The coders were able to arrive at similar categorizations, regardless of the school subject concerned (Table 2).

In the group of reliability coefficients indicative of agreement between pairs of coders, statistically significant differences were in evidence (Tables 2 and 3). The reliability coefficients for the pairs including Observer 2 were systematically lower as compared with the others. This observer's conception

of the categorizing directions had differed systematically from the other observers' conceptions. The coefficients indicating agreement between Observers 1, 2 and 4 represented a rather satisfactory level.

Table 2. The Pi Coefficients Computed from the Material

	Type of coefficient	$\bar{X}$	s	N
1.	Total sample	.79	.08	84
2.	Inter-coder agreement T1	.79	.09	60
3.	Inter-coder agreement T2	.80	.06	24
4.	Between-coder constancy	.71	.12	48
5.	Within-coder constancy	.74	.13	16
6.	Coder pair 1, 2	.74	.10	14
7.	Coder pair 1, 3	.84	.03	14
8.	Coder pair 1, 4	.83	.03	14
9.	Coder pair 2, 3	.73	.08	14
10.	Coder pair 2, 4	.76	.09	14
11.	Coder pair 3, 4	.83	.04	14
12.	Religion	.81	.06	24
13.	Civics	.77	.09	12
14.	Arithmetic	.76	.12	12
15.	Finnish	.80	.08	12

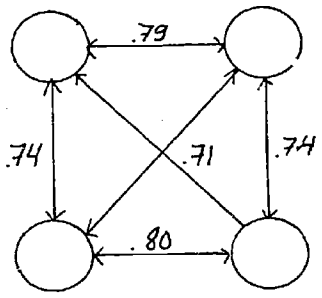
Table 3. Significance of the Differences between Coder Pair Reliabilities, t test

	Coder pair					
	1,2	1,3	1,4	2,3	2,4	3,4
1,2		-3.89	-3.73	-0.42	0.47	-3.27
1,3			0.30	4.71	3.11	0.61
1,4				4.61	3.00	0.35
2,3					-0.91	-4.20
2,4						2.70

____ = p .01

In interpreting the table, the effect of overlapping classifications on the risk-level limits should be taken into consideration (see Hays 1963, 375-376; 471-472; 483-485).

The coefficients for the first coding did not differ significantly from those for the second coding. Inter-coder agreement was .79 during the autumn term (T1) and .80 during the spring term (T2). When the two coding occasions were compared so as to determine the between-coder constancy, highly significant differences were found (T1/comparison, $t = 3.72$, $df = 121$, $p < .001$; and T2/comparison, $t = 3.85$, $df = 83$, $p < .001$). Inter-coder agreement was high on both occasions, whereas between-coder constancy was rather poor. This state of affairs can be illustrated by the following graphic representation.



There is reason to assume that the codings performed during the autumn term, when the time lapse since coder training proper was comparatively short, were better estimates of "correct" codings than were the spring term codings, which had changed for all the coders in the same direction.

Within-coder constancy was slightly higher in comparison with between-coder constancy, though not to a statistically significant extent. The result supports the interpretation that the coder group as a whole shifted in the same direction in the use of the categorizing criteria between the points of time T1 and T2.

5. An Appraisal of the Overall Reliability Results

The results obtained concerning overall reliability suggest that the alterations made in Flanders's original categories did not worsen agreement, at least not decisively, even though the number of erroneous categorizing possibilities increased and the degree of agreement to be expected on the basis of chance diminished. The coefficients obtained here can be compared with

reliabilities obtained in other studies. Hough, Lohman & Ober state that if Scott's pi equals .60, this can be regarded as a minimum proficiency (1969). According to Flanders, "a Scott coefficient of .85 or higher is a reasonable level of performance" (1967, 166). The average coefficient obtained in the present study was slightly lower. As a general rule, however, the reliability coefficients obtained in previous studies have not quite reached the limit suggested by Flanders (Hough & Ober 1967, 334).

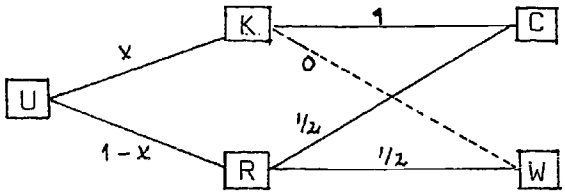
The change that was found to take place in the principles of categorization, judging by both the between-coder and within-coder constancy comparisons, has relevance to coder training as well as to the treatment of the material. Agreement controls carried out at given intervals are not enough to avoid systematical errors in coding; on the contrary, constancy control through time must also be resorted to. It is generally inadvisable to code any material in chronological order, since trends due to the observer's behaviour may then be shown by the measurements. The emergence of such trends can effectively be prevented by randomizing the order of coding.

The estimation of overall reliability rests on the assumption that the units assigned by various coders to a category are the same if they are equal in number. Several studies speak, however, against this assumption. Observers may commit mistakes offsetting one another, with the result that the marginal distributions will remain identical. Scott's pi will then yield excessively high values. In studies where reliability has been examined unit by unit, agreement coefficients 10 to 20 percent lower in value have been obtained on an average (Waxler & Mishler 1966). It sounds paradoxical that, where two coders do the coding completely at random, the pi coefficient will approach unity, yet this is the case since the marginal distributions will eventually become similar in shape. Also, the shift phenomenon observed in the use of the categorizing principles merits closer examination. Thus, the overall reliability method

must be supplemented by a method through which the reliability of any one individual category can be determined.

6. The Single-unit Single-category Coding Situation

When the coder assigns any one unit to a dichotomous category, this can be described, from the viewpoint of probability theory, in the following way (Schutz 1952, 121-122; Guetzkow 1950, 51):



U = unit to be coded
K = coding according to criterion
R = random coding
C = "correct" coding
W = "incorrect" coding

Thus, categorizing may turn out correctly either because the coder uses the criterion of by chance. The probability for the coder's employing the criterion can be computed. What we need to find is the probability with which the unit U is coded correctly by employing the criterion K. From Bayes's rule we have

$$(3) \qquad p_{k,c} = \frac{2x}{x + 1}$$

Yet we do not know the value of x. What we can observe is only agreement (A), which also includes the correct coding due to chance factors ($p_{r,c}$). Here, A is the probability for the coding

to be correct (p_c).[1] From the rule of elimination,

$$p_c = \frac{1}{2}(x + 1) \quad (4)$$

and thus,

$$x = 2A - 1 \quad (5)$$

Substituting x into (3),

$$p_{k,c} = \frac{2A - 1}{A} \quad (6)$$

whence

$$A = \frac{1}{2 - p_{k,c}} \quad (7)$$

Now a value can be computed for empirical agreement (A) such that the probability with which the coder employs the criterion in arriving at a correct result will be, say $p_{r,c} = .90$. The value of A will then be .91. The matter is not invariably so simple in practice. Above we assumed that the correct coding was known. Agreement (A) between two coders may, however, also be due to the fact that both categorize the same unit in the same way but incorrectly. Who does, in the last resort, decide the correctness of coding? In other words, who can be said to proceed in a perfectly reliable manner in employing the criterion? The absence of an ultimate criterion, in combination with the fact that the category systems are rarely dichotomous, complicates the reliability analysis concerning the individual category.

---

[1] Here, p_c is the theoretical and A the observed value.

On the basis of the above argument and one of Osgood's content analysis models (Osgood 1959, 33-88), the present writer developed a method for the estimation of reliability values for each of the thirteen categories used. The following requirements were imposed on the method:

1) Each category was to be dichotomized, so that use could be made of the above probability model.

2) The unit was to be such that each of the two coders would actually base his categorizing on the same unit.

3) All the categories were to be considered simultaneously, to ensure that all the reliability results would rest on activities comparable to original coding and that they would not be unrealistically high.

The following two assumptions were made:

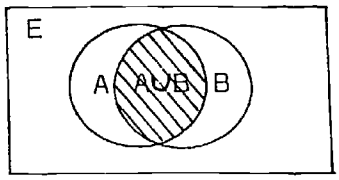
1) Instances of reliable coding are those where two independently acting coders show that they have simultaneously perceived the occurrence of a behaviour belonging to a given category. Perception of the absence of a behaviour is not regarded as reliable categorizing.

2) The true frequency of events in a category is the mean of the frequencies observed by the coders. The correct frequency is unknown, but the mean is supposed to be the best available estimate of it (Wright 1967, 96).

The method will be described below as it is applied to any one dichotomized category. In this study, all the categories were on the same coding sheet (Appendix 2), and the categorizing principle was exactly the same for each of them. A and B are two coders whose agreement about one category is in question. At a point indicated in the videotape, the seconds counter is started and coding begins. During the first ten seconds the coders have to observe the train of events on the videotape. During the following ten seconds they do not attend to the videotape but, instead, indicate on the coding sheet whether activities falling in the category concerned was in evidence during the

preceding ten-second period. Ten-second periods of observation and note-taking alternate for some 30 minutes. The situation concerning any one category can be represented as in Figure 8 (see Kerlinger 1964, 67-80).

Figure 8. A Schematic Representation of Reliability for One Category



E = basic set or the set of all observation periods

A = the subset of E consisting of those units that, according to coder A, included behaviours satisfying the categorizing criterion

B = the subset of E consisting of those units that, according to coder B, included behaviours satisfying the categorizing criterion

C = A ∩ B = the subset of E coded in a reliable manner, according to the assumption

N_E = the number of units (i.e., observation periods) in the basic set

N_A = the number of units in subset A

N_B = the number of units in subset B

N_C = the number of units in subset C

$\frac{1}{2}(N_A + N_B)$ = on the assumption made, the best estimate of the frequency in the basic set of the behaviours belonging to the category

Agreement for the category is obtained from

$$A = \frac{N_C}{\frac{1}{2}(N_A + N_B)} \quad (7)$$

indicating what proportion of the estimated number of units belonging to the category was coded in a reliable manner. The agreement due to chance can be computed from the formula

$$A_r = \frac{N_A \cdot N_B}{N_E^2} \quad (8)$$

Preliminary experiments showed that the 10-second observation period was suitable. When use was made of all 13 categories, each unit period generally included 1 - 4 behaviours representing the various categories.

7. Results concerning the Reliabilities of the Various Individual Categories

A total of 2 072 agreement coefficients (A) were computed, each of which rested on some 90 observation periods. The corresponding expectations (A_r) were also computed, but the writer feels that it is unnecessary to report them here. In each case the agreement coefficients exceeded their expected values very definitely and to a statistically significant degree (Table 4).

The coefficients were examined by means of one-way analysis of variance. This method was chosen because the groups to be compared were usually more than two in number. A total of 112 analyses of variance were computed, 56 of which (those for the A coefficients) are presented in this report. Such a large number of statistical testing is likely to include cases where statistical significance is due to chance. Moreover, the consecutive analyses are not mutually independent, and thus the prob-

ability of the rejection error will exceed the chosen risk level. F values for which p > .01 were not regarded as significant.

There were considerable differences between the categories in reliability. Categories 3, 4b and 9b seemed the poorest in this respect (Table 4). Nevertheless, the values for all the categories definitely exceeded the corresponding mathematical expectations. No standard deviations are given in the following tables, as these were, by and large, equal to those set out in Table 4. As was to be expected, "pupil answers" (8) and "teacher lectures" (5) were the two most clear-cut categories, judging by both the high coefficients and the standard deviations.

Consideration of the coefficients by school subjects reveals a number of significant differences (Table 5). Arithmetic lessons had apparently been the most difficult to code. This was perhaps due to the general nature of these lessons, which contain a lot parallel and intertwined interaction, associated with blackboard work, etc. Individual guidance and blackboard work presented particular interpretational difficulties in coding. On the other hand, religion lessons seemed to be the easiest to code, judging by this material.

Analysis of the differences between coder pairs supported the finding made in considering the overall reliability that Observer 2 had done the coding differently from the other three (Table 6). The coefficients for the pairs including Observer 2 were all lower in comparison with the rest of the coefficients. The other coders had carried out the codings more uniformly and no clear differences were perceptible between them.

Regarding the coding occasions the following was observed. Inter-coder agreement diminished with time (Table 7). This was so for almost all categories. The only exception was provided by "pupil answers" (8), in the case of which the categorizing had remained more or less unchanged.

Table 4. The Agreement Coefficients (A) as Computed from the Data and their Corresponding Expectations (A_r)

ategory		1	2	3	4a	4b	5	6	7	8	9a	9b	10	Z	$\bar{X}$
gree-	$\bar{X}$	.77	.62	.58	.70	.54	.79	.73	.76	.88	.70	.52	.73	.77	.67
ent (A)	s	.13	.29	.25	.21	.29	.12	.17	.16	.13	.16	.31	.15	.26	.10
xpec-	$\bar{X}$	.04	.00	.01	.04	.01	.14	.05	.02	.12	.02	.00	.11	.01	.04
ation (A_r)	s	.03	.00	.00	.03	.01	.08	.02	.02	.08	.01	.00	.11	.00	.01

= 148

Table 5. The Agreement Coefficients, by School Subjects

ategory	1	2	3	4a	4b	5	6	7	8	9a	9b	10	Z	$\bar{X}$
eligion	.84	.71	.65	.60	.65	.89	.69	.76	.92	.75	.52	.74	.87	.73
ivics	.77	.58	.59	.79	.51	.78	.71	.77	.86	.67	.54	.61	.63	.65
rithmetic	.70	.37	.53	.86	.42	.74	.81	.72	.76	.70	.65	.88	-	.61
innish	.71	.77	.48	.60	.48	.67	.76	.79	.93	.62	.33	.70	.68	.65
	7.38	9.01	1.88	10.30	2.51	28.77	1.96	0.65	9.13	3.04	2.96	15.03	5.23	8.13
f 1	3	3	3	3	3	3	3	3	3	3	3	3	2	3
f 2	80	76	76	80	74	80	80	80	80	80	77	80	53	80
<	.01	.01		.01		.01			.01			.01	.01	.01

Table 6. The Agreement Coefficients, by Coder Pairs

Category		1	2	3	4a	4b	5	6	7	8	9a	9b	10	Z	$\bar{X}$
Coder pair	1, 2	.72	.56	.55	.63	.48	.73	.64	.67	.87	.70	.42	.67	.68	.61
	1, 3	.87	.69	.68	.79	.65	.84	.79	.88	.92	.76	.62	.80	.86	.74
	1, 4	.78	.72	.60	.78	.61	.82	.80	.82	.90	.74	.65	.81	.92	.73
	2, 3	.73	.58	.61	.64	.50	.76	.68	.70	.82	.62	.42	.69	.68	.62
	2, 4	.72	.59	.46	.57	.43	.74	.70	.66	.85	.65	.42	.66	.68	.61
	3, 4	.80	.62	.57	.77	.58	.84	.79	.83	.90	.72	.60	.77	.85	.71
F		3.07	0.66	1.19	3.09	1.12	3.12	2.41	6.28	1.20	1.46	1.76	3.33	1.90	7.21
df 1		5	5	5	5	5	5	5	5	5	5	5	5	5	5
df 2		78	74	74	78	72	78	78	78	78	78	75	78	50	78
p <									.01				.01		.01

Table 7. The Agreement Coefficients and Constancy Coefficients (Comparison between Coding Occasions T1 and T2)

Category	1	2	3	4a	4b	5	6	7	8	9a	9b	10	Z	$\bar{X}$
Inter-coder T1	.79	.67	.58	.70	.62	.80	.76	.81	.88	.75	.63	.75	.80	.70
Inter-coder T2	.72	.52	.58	.69	.37	.75	.66	.63	.86	.55	.24	.70	.66	.58
Between-coder T1/T2	.66	.49	.44	.67	.40	.69	.62	.67	.88	.59	.30	.63	.72	.57
Within-coder T1/T2	.66	.53	.50	.72	.44	.72	.67	.69	.88	.64	.34	.62	.71	.60
F	13.88	5.54	4.06	0.07	11.15	8.91	10.76	20.34	0.61	23.99	26.74	8.63	1.67	35.55
df 1	3	3	3	3	3	3	3	3	3	3	3	3	3	3
df 2	145	141	141	145	139	145	145	145	145	145	138	145	89	145
p <	.01	.01	.01		.01	.01	.01	.01		.01	.01	.01		.01

Table 8. The Agreement Coefficients and the Order of Coding

Category	1	2	3	4a	4b	5	6	7	8	9a	9b	10	Z	$\bar{X}$
Coded first	.75	.54	.58	.72	.56	.77	.73	.75	.87	.67	.42	.77	.77	.65
Coded second	.80	.72	.57	.67	.52	.82	.73	.77	.88	.73	.64	.68	.77	.70
F	3.75	8.42	0.01	1.24	0.41	3.11	0.00	0.43	0.23	2.44	10.32	8.06	0.00	4.89
df 1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
df 2	82	78	78	82	76	82	82	82	82	82	79	82	54	82
p <		.01									.01	.01		

As was to be expected between-coder constancy was slightly lower than within-coder constancy. Only in categories 4b, 8 and Z were differences not significant. No consistent shift, comparable to that discovered in the context of overall reliability was here in evidence. The changes that had taken place between T1 and T2 varied in direction, depending on the coder. The changes differed from one category to another, and no changes had occured for categories 8 and 4a. These two categories in fact represent forms of behaviour that comparatively infrequently necessitate an interpretation of the total situation on the part of the coder.

The codings were performed during two consecutive sessions of an hour's duration on each of the coding days. An analysis of variance concerning the order of coding was possible to compute. No fatique effect was in evidence (Table 8). As a matter of fact, the second codings were even better.

The overall reliability method and the determination of reliability by individual categories provided a similar picture of the observers' coding proficiency. The computed probability (p_c) with which a unit is correctly categorized may be used as a measure of the observer's accuracy. Provided that the inter-coder agreement coefficients for two or more coders are known, the coders' accuracy can be estimated (see Guetzkow 1950, 54 and Bernstein 1969, 49-52).

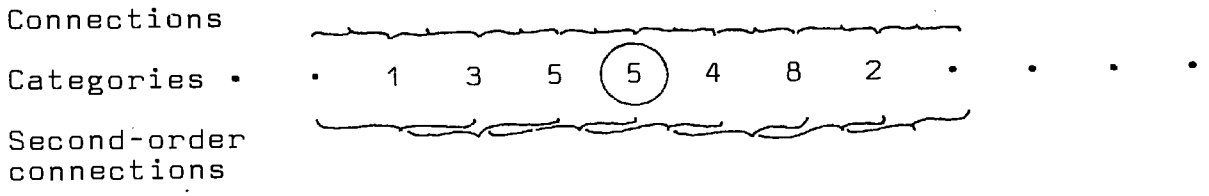
The accuracy coefficients computed from the two types of reliability estimates were largely similar (Table 9).

Table 9. Coder Accuracy (1 = as computed from overall reliabilities; 2 = as computed from the mean yielded by the method developed by the writer)

	Method of computation of agreement	
	1	2
Coder 1	.92	.87
Coder 2	.80	.71
Coder 3	.91	.86
Coder 4	.91	.84

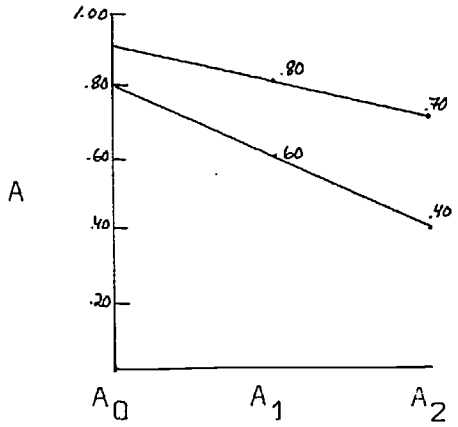
8. Consideration of the Results about Reliability

Waxler & Mishler state that "no simple prescription can be offered either about how to compute an index of reliability or what level to set acceptable" (1966). Unfortunately, investigators have generally been content with ascertaining the presence of some sort of reliability, without troubling themselves with the question of what kinds of treatment and operations are justified by it. It is obvious that in an intensive case study a comparatively high level of reliability is a necessity. Another point should also be taken into consideration in interactional research. What I have in mind can be demonstrated by brief example. Let us assume that the coder of a given material has obtained the following sequence of numbers, the encircled number representing erroneous coding.



In the estimation of overall reliability, erroneous coding causes a change in the profile and slightly reduces agreement (A_c). When the sequence of numbers is tabulated by connections into the interaction matrix, two of the connections are found to fall in incorrect cells (A_1). In a second-order Markov chain, three sequences of events, i.e., 3-5-5, 5-5-4 and 5-4-8, will be misplaced (A_2). It can be shown that agreement declines fast when we proceed to higher-order connections. How fast the decline is depends on the starting level.

Figure 1. The Decline in Agreement in Shifting from the Profile to Connections



It will be seen that the starting level (i.e., inter-coder agreement) should exceed .80; since if it is lower, the second-order Markov chains will make little sense and the interaction matrix becomes undependable.

One qualification is necessary here, however: there are various sorts of error, although no attention to the meaningfullness of errors was paid in this study.

However, perfectionism with regard to reliability is not likely to be an appropriate goal. Brown & Webb hit the mark in stating: "A team of observers can be trained to the point of near-perfect agreement, but this does not eliminate the possibility that instead of making numerous subjective judgments of a differing and conflicting nature (as they did prior to learning), they now make only one - the same one" (1968, 35). The rules guiding the coders must not be a means to attain a high reliability; instead, they should be a method intended to facilitate the measurement of a theoretically important concept.

References:

BALES, R. F. 1950. Interaction Process Analysis: A Method for the Study of Small Groups. Cambridge, Mass.: Addison & Wesley

BERELSON, B. 1954. "Content Analysis". Handbook of Social Psychology, Volume I (Ed. LINDZEY, G.) Cambridge, Mass.: Addison & Wesley, 488-522

BERNSTEIN, A. L. 1969. "An Estimate of the Accuracy of Nominal Category Coding". Classroom Interaction Newsletter 4, 49-52

BORGATTA, E. F. & BALES, R. F. 1953. "The Consistency of Subject Behavior and the Reliability of Scoring in Interaction Process Analysis". American Sociological Review 18, 566-569

BROWN, B. B. & WEBB, J. N. 1968. "Valid and Reliable Observations of Classroom Behavior". Classroom Interaction Newsletter 4, 35-38

COHEN, J. A. 1960. "A Coefficient of Agreement for Nominal Scales" Educational and Psychological Measurement 20, 37-46

FLANDERS, N. A. 1965. Teacher Influence, Pupil Attitudes, and Achievement. Washington D.C.: U.S. Government Printing Office

FLANDERS, N. A. 1967. "The Problems of Observer Training and Reliability". Interaction Analysis: Theory, Research and Application (Eds. AMIOON, E. J. & HOUGH, J. B.) Palo Alto: Addison & Wesley, 158-166

GUETZKOW, H. 1950. "Unitizing and Categorizing Problems in Coding Qualitative Data". Journal of Clinical Psychology 6, 47-58

HAYS, W. L. 1963. Statistics for Psychologists. New York: Holt, Rinehart & Winston

HOUGH, J. B. & LOHMAN, E. E. & OBER, R. 1969. "Shaping and Predicting Verbal Teaching Behavior in a General Methods Course". Journal of Teacher Education 20, 213-224

HOUGH, J. B. & OBER, R. 1967. "The Effect of Training in Interaction Analysis on the Verbal Teaching Behaviors of Pre-Service Teachers". Interaction Analysis: Theory, Research and Application (Eds. AMIDON, E. J. & HOUGH, J. B.) Palo Alto: Addison & Wesley, 329-346

KERLINGER, F. N. 1964. Foundations of Behavioral Research: Educational and Psychological Inquiry. New York: Holt, Rinehart & Winston

KOGAN, L. S. & HUNT, V. 1950. "Problems of Multi-Judge Reliability". Journal of Clinical Psychology 6, 16-19

KOSKENNIEMI, M. & KOMULAINEN, E. & AL. 1969. "Investigations into the Instructional Process. I. Some Methodological Problems". Institute of Education University of Helsinki Research Bulletin No 26

MEOLEY, O. M. & MITZEL, H. E. 1963. "Measuring Classroom Behavior by Systematic Observation". Handbook of Research on Teaching (Ed. GAGE, N. L.) Chicago: Rand McNally 247-328

OSGOOD, C. E. 1959. "The Representational Model and Relevant Research Methods". Trends in Content Analysis (Ed. POOL, I.) Urbana: University of Illinois Press, 33-88

SCHUTZ, W. C. 1952. "Reliability, Ambiguity and Content Analysis". Psychological Review 59, 119-129

SCOTT, W. A. 1955. "Reliability of Content Analysis: The Case of Nominal Scale Coding". Public Opinion Quarterly 19, 321-325

STUKÅT, K-G. 1966. Pedagogisk Forskningsmetodik. Stockholm: Almqwist & Wiksell

WAXLER, N. E. & MISHLER, E. G. 1966. "Scoring and Reliability Problems in Interaction Process Analysis: A Methodological Note". Sociometry 29, 28-40

WRIGHT, H. F. Recording and Analyzing Child Behavior. New York: Harper & Row

Appendix 1.

The Employed Classification System

Teacher talk	1 .	Accepts, praises or encourages
	2 .	Corrective feedback
	3 .	Uses pupil ideas
	4a.	Asks narrow questions
	4b.	Asks broad questions
	5 .	Expresses information or own opinions
	6 .	Gives directions
	7 .	Criticizes pupil behaviour
Pupil talk	8 .	Answers to a question
	9a.	Relevant spontaneous talk and suggestions
	9b.	Irrelevant spontaneous talk
Others	10 .	Silent work, individual work or guidance
	Z .	Tumult, confused situation

Appendix 2.

The Coding Sheet Employed in Estimating the Reliabilities of the Individual Categories

Unit No.	1	2	3	4a	4b	5	6	7	8	9a	9b	10	Z
1													
2													
3													
4													
etc.													

Coder: _____

Date: _____ Time: _____

Lesson: _____

Date and time of rec: _____

_____

Starting point: _____

_____

Previous Issues in this Series, Available from the Institute

No. 17 The Finnish Senior Secondary Research Project I. General Presentation of the Project and Its Original Subjects by ANNA-LIISA SYSIHARJU. Sept., 1967. 31 pp.

No. 18 The Finnish Senior Secondary Research Project II. Departure from Junior Secondary and Transfer to Senior Secondary by ANNA-LIISA SYSIHARJU. Oct., 1967. 52 pp. Out of print.

No. 19 Relationships between Absences from School and Certain Factors Related to School. A Study on Elementary School Grades Three to Six by PERTTI KANSANEN. Nov., 1967. 10 pp.

No. 20 On the Anxiety Associated with School Attendance and the Grammar School Entrance Examination by EEVA PATJAS. May, 1968. 16 pp.

No. 21 School Achievement and Personality. I: Design and Hypotheses, IV: Results and Discussion by ERKKI A. NISKANEN. Oct., 1968. 51 pp.

No. 22 School Achievement and Personality. II: Operations at the Variable Level by ERKKI A. NISKANEN. Oct., 1968. 124 pp.

No. 23 School Achievement and Personality. III: Operations at the Factor Level by ERKKI A. NISKANEN. Oct., 1968. 39 pp.

No. 24 The Effectiveness of Two Methods of Teaching English as a Foreign Language in Some Finnish Secondary Schools by DANIEL J. CASEY. Nov., 1968. 37 pp.

No. 25 The Learning of Elementary Algebra. An Empirical Investigation of the Results of Learning in a Simplified School Learning System by PAAVO MALINEN. Jan., 1969. 85 pp.

No. 26 Investigations into the Instructional Process. I. Some Methodological Problems by MATTI KOSKENNIEMI and ERKKI KOMULAINEN in collaboration with Anna-Kaarina Falck and Pentti Holopainen. Oct., 1969. 35 pp.

Distribution by EDUCA r.y., Snellmaninkatu 10 A, Helsinki 17